1. For a given real vector with elements $z = [z_1, \ldots, z_n]^T$, let the function $s : \mathbb{R}^n \to \mathbb{R}^n$ be,

$$s(z) = \frac{1}{\sum_{i=1}^n e^{z_i}} [e^{z_1} \quad \ldots \quad e^{z_n}]^T.$$

This is sometimes called the *softmax* function. Establish the following properties of $s(\cdot)$:

(a) $s(z) = s(z + c\mathbf{1})$ where $\mathbf{1}$ is a vector of 1's and c is a constant.

(b) For any vector z, $s(z)$ is a probability vector. That is,

$$\sum_{i=1}^n s(z)_i = 1, \qquad \text{and} \qquad s(z)_i \in [0, 1].$$

(c) Take now a sequence of vectors $z^{(1)}, z^{(2)}, \ldots$ that converges to a limit,

$$\lim_{t\to\infty} z^{(t)} = z^*,$$

with $z^*_{i^*} > z^*_j$ for all $j \neq i^*$. Further consider the sequence $s(z^{(1)}), s(z^{(2)}), \ldots$. Establish that,

$$\lim_{t\to\infty} s(z^{(t)}) = s(z^*),$$

and that $s(z^*)_{i^*}$ is the *mode* of this limiting probability distribution.

2. Let the function $\sigma(u) = max(u, 0)$. This is sometimes called the *ReLU activation function*. Consider the matrices $A_1 \in \mathbb{R}^{3\times 2}$ and $A_2 \in \mathbb{R}^{2\times 1}$, and the vectors $b_1 \in \mathbb{R}^2$ and $b_2 \in \mathbb{R}$. Consider now the function $\rho : \mathbb{R}^3 \to \mathbb{R}$ constructed via,

$$\rho(x) = \sigma(A_2\sigma(A_1 x + b_1) + b_2),$$

where $\sigma(\cdot)$ applied to a vector is applied element-wise.

(a) Assume b_1 and b_2 are zero vectors and that A_1 and A_2 have all entries 1. Determine $\rho([1 \;\; 1 \;\; 1]^T)$.

(b) Assume A_1 has first two rows as the 2×2 identity matrix and third row with zero entries. Further assume $b_1 = [0 \;\; 0]^T$, $A_2 = [1 \;\; 0]^T$, and $b_2 = 10$. For what values of x do we have $\rho(x) \neq 0$?

(c) It can be shown that for any values of A_1, A_2, b_1, and b_2 we can represent $\rho(\cdot)$ as,

$$\rho(x) = C_{i(x)}\; x + d_{i(x)},$$

where C_i and d_i are fixed for each $i = 1, \ldots, K$ and $i(x) : \mathbb{R}^3 \to \{1, \ldots, K\}$ determines a *piece-wise affine* partition of $\mathbb{R}^3$. That is, it partitions $\mathbb{R}^3$ into K polytopes and returns the index of the associated polytope. Determine an upper bound on K.

3. We consider now the least squares problem, where we attempt to solve the linear system of equations $Aw = b$, with A an $m \times n$ matrix and b an m-vector. Consider the loss function,

$$L(w) = ||Aw - b||^2,$$

where the norm is the L_2 norm and we are seeking w that minimizes $L(w)$.

   (a) Explain why the minimizer of $L(w)$ is the same as the minimizer of $||Aw - b||$.

   (b) Show that the gradient of $L(\cdot)$ is $\nabla L(w) = 2A^T(Aw - b)$.

   (c) Establish that $L(w)$ has a unique minimizer if and only if the columns of A are linearly independent.

4. Consider a sequence of i.i.d. (independent and identically distributed) uniform$(0, 1)$ random variables denoted $U_1, U_2, \ldots$.

   (a) Define the sequence of random variables $Y_1, Y_2, \ldots$ via,

$$Y_i = \frac{U_1 - a}{b},$$

   for constants a and $b > 0$. Find a and b such that the expectation of Y_1 is 0 and the variance is 1.

   (b) Consider now sequence, $\overline{X}_1, \overline{X}_2, \ldots$, where

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} U_i.$$

   You are interested in the probability

$$\mathbb{P}\Big(2\sqrt{3n}\left(|\overline{X}_n| - 0.5\right) > 2\Big).$$

   Use the central limit theorem to approximate this probability for non-small n.

   (c) Estimate the above probability for $n = 3$ using Monte-Carlo by generating a million instances of $\overline{X}_5$ and compare to the normal approximation above. Repeat for $n = 30$ and compare again. Use the programming language of your choice (R, Python, Matlab, Julia, etc...).

5. Write programs or script using the programming language of your choice that perform the following tasks. Use programming primitives as opposed to dedicated library functions:

   (a) Given an input matrix, check if it is: Symmetric, upper-triangular, lower-triangular.

   (b) Given a list of numbers, return a sorted list of numbers.

   (c) Given a sequence of n input vectors, each of length p, compute the $p \times p$ sample covariance matrix of these vectors.

   (d) Use brute-force Riemann sums to illustrate numerically that

$$\int_{x_1=-5}^{5}\int_{x_2=-5}^{5} \frac{1}{2\pi} e^{-\frac{x_1^2+x_2^2}{2}}\, dx_1\, dx_2 \approx 1.$$